## Chapter 19:

# Spatial Regression Modeling[1]

**Ned Levine**
Ned Levine & Associates
Houston, TX

**Dominique Lord**
Zachry Dept. of Civil Engineering
Texas A & M University
College Station, TX

**Byung-Jung Park**
Korea Transport Institute
Goyang, South Korea

**Srinivas Geedipally**
Texas Transportation Institute
Arlington, TX

**Haiyan Teng**
Houston, TX

**Li Sheng**
Houston, TX

[1] This chapter was the result of the efforts of several people. Dr. Shaw-pin Miaou of College Station, TX designed the MCMC algorithm for the Poisson-Gamma-CAR model. Dr. Byung-Jung Park modified the algorithm to incorporate Poisson-Gamma-SAR, Poisson-Lognormal-CAR/SAR, and the MCMC binomial-CAR/SAR models. Dr. Srinivas Geeidpally added the MCMC Normal-CAR/SAR models. Dr. Dominique Lord of Texas A & M University provided technical consulting on the dispersion parameters in these models. Dr. Ned Levine developed the block sampling scheme and provided overall project management. Ms. Haiyan Teng and Dr. Li Sheng programmed the routines and added numerous technical improvements to the algorithms. The authors thank Dr. Richard Block for testing the routines.

# Table of Contents

# Chapter 19:
# Spatial Regression Modeling

In this chapter, we examine spatial regression modeling using the the Markov Chain Monte Carlo (MCMC) method. Users should be thoroughly familiar with the materials in Chapters 15, 16, 17 and 18 before attempting to read this chapter. A good background in statistics is necessary to understand the material.

## Spatial Regression Modeling

Spatial regression involves adding a spatial component into a regression model. There are two major ways to express this component, either as an explicit spatial variable or as an internally-estimated spatial parameter. There are advantages and disadvantages to each approach and frequently they are included together.

### Explicit Spatial Variable

With an explicit spatial variable, a specific spatial relationship is added as an independent variable. Examples of this are the area of the zone, the distance to the central city, the distance to a particular facility, or an average value of the dependent variable for nearby zones.

The justification for including an explicit variable depends on what is being modeled. For instance, spatial statisticians frequently distinguish between *global* and *local* effects. Global effects are those that cover the entire study region whereas local effects affect only a small geographic area. Without distinguishing those two types of effects, ambiguity can be produced in a model.

#### *Area of the zone*

One of the most well known spatial variables that should be included in any statistical model is the area of the zone. Typically, zones based on a census will have a size that is proportional to their residential population. Thus, zones in the center of a metropolitan area will typically be very small, perhaps single blocks, while zones in the suburbs will be very large, covering several square miles. Without adjusting for the size of the zone, distortions in estimates can be produced. For example, all other things being equal, more events can occur within a larger zone than for a smaller zone. Modelers will frequently include the area of the zone as a statistical control variable.

Note that with zones based on census geography (e.g., census tracts, traffic analysis zones), there will be a negative correlation between the area of a zone and its distance from the metropolitan center (see below). With census-based geography zones, get bigger with distance from the metropolitan center. Thus, the analyst must be aware of this while interpreting regression coefficients.

### *Shape of the zone*

Related to this is the shape of the zone. If two zones have very different shapes (e.g., one is square while the other is pointed and long and narrow), allocation error (and, hence, modeling error) is liable to be greater in the one that is more irregular, all other things being equal, than in the one that is square. This is the so-called *Modifiable Area Unit Problem* (or MAUP) problem (see Wikipedia, 2012; Hipp, 2007; Wooldridge, 2002; Openshaw, 1984).

There is not a simple statistical variable that can be included to adjust for irregularity, short of some fractal measure (Lam & De Cola, 1993). Ideally, if the zones could be uniform grid cells, then distortions due to shape can be minimized. Otherwise, the user needs to be cognizant of the potential for shape to influence the coefficients of a model and be prepared to modify the data to incorporate irregular boundaries (e.g., smoothing the distribution of events in a hot spot that are assigned to large zones to reduce shape effects; see Chapter 11 on Head-Bang Interpolation).

### *Distance from downtown*

Another well known spatial variable that should be included in a spatial regression model is the distance from the zone to the central area in the study region. For example, with data from a city or metropolitan area, this variable would be the distance (in miles or kilometers) to the downtown area. Typically, the density of events is a function of distance from the central city primarily due to land costs. This effect has been studied as far back as the early 19$^{th}$ century with the work of von Thünen (1826). Alonso (1964) modernized the framework by demonstrating that each activity has its own land price (i.e., the cost of the land underlying the activity) and that a spatial equilibrium will be established in terms of the relative price of different activities.

All other factors being equal, there will be more events occurring in the center of a metropolitan area than in the periphery primarily due to the increased concentration of activities (which is a function of the underlying land costs). Frequently, there will be a relationship between distance from the downtown area and a variable of interest. For example, Levine (2011) showed that the risk of motor vehicle crashes was double in downtown Houston than in the suburbs. This was a function of the concentrated traffic in the downtown area, the greater

number of intersections that created potential conflicts between drivers, and the greater number of driveways. Further, male drivers were more likely to be involved in a crash in the downtown area than female drivers so that part of the increased risk was due to a predominance of male drivers.

In another study, Levine and Lee (2013) showed that the distance traveled for crime trips (journey-to-crime) was associated with the distance an offender lived from central Manchester, England with a negative binomial model. Further, different types of crime were associated with specific travel distances, with property crimes being much longer, on average, than violent crime. Further, these distances were mediated by the distance the offender lived from the city center. Around one-fifth of the crimes committed by females were shoplifting and these were much more likely to occur in the city center or in one of the suburban town centers.

To see this, Figure 19.1 shows an estimate of the number of crimes committed in the City of Houston from 2007 to 2009 by distance from downtown Houston in quarter mile intervals. The data were 807,788 reported crime incidents and the estimate was produced by the *CrimeStat* journey-to-crime interpolation routine using a normal kernel and an adaptive bandwidth with a minimum of 200 crimes (see Chapter 13). As seen, the number of crimes per quarter mile increases from about 3,000 in downtown Houston to more than 20,000 at about 11 miles from downtown Houston. The number of crimes then drops rapidly, primarily because the search radius extends beyond the boundaries of the City of Houston.

However, each quarter mile 'band' covers a larger area. Consequently, one would expect, all other things being equal, for there to be more events with distance from downtown. It is essential to normalize this measure to allow equal comparisons. Consequently, we divided the number of crimes per quarter mile band by the area (in square miles) covered by each band.

Figure 19.2 shows the results. As seen, the number of crimes per square mile is greater than 30,000 in downtown Houston but drops very dramatically with distance. The curve is almost a perfect negative exponential function and is frequently modeled by that function (see Chapters 13 and 28). This could be converted into a probability estimate by dividing each density by the total density of all bins. In other words, the probability of a crime being committed in downtown Houston decreases rapidly with distance from downtown.

The point is that one should include explicit spatial variables to account for these potential global effects, if only for statistical control. Including a global spatial variable such as the distance from downtown makes explicit how the dependent variable changes by distance, such as in the crash risk study mentioned above (Levine, 2011). As was discussed in Chapter 6, local spatial autocorrelation is frequently a function of global spatial effects. Typically, the closer to the center of the city a zone is, the more likely that there will be correlations between

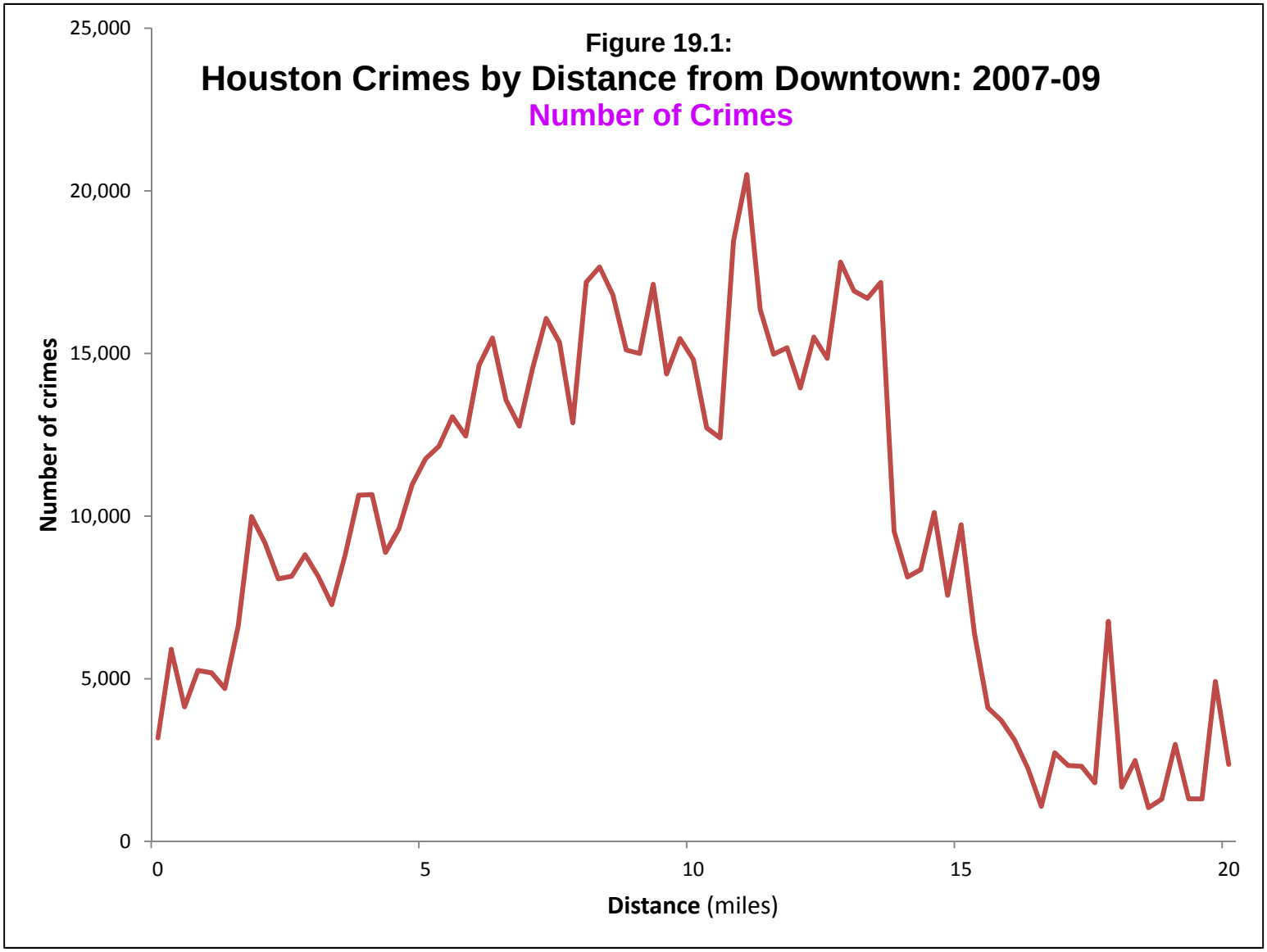
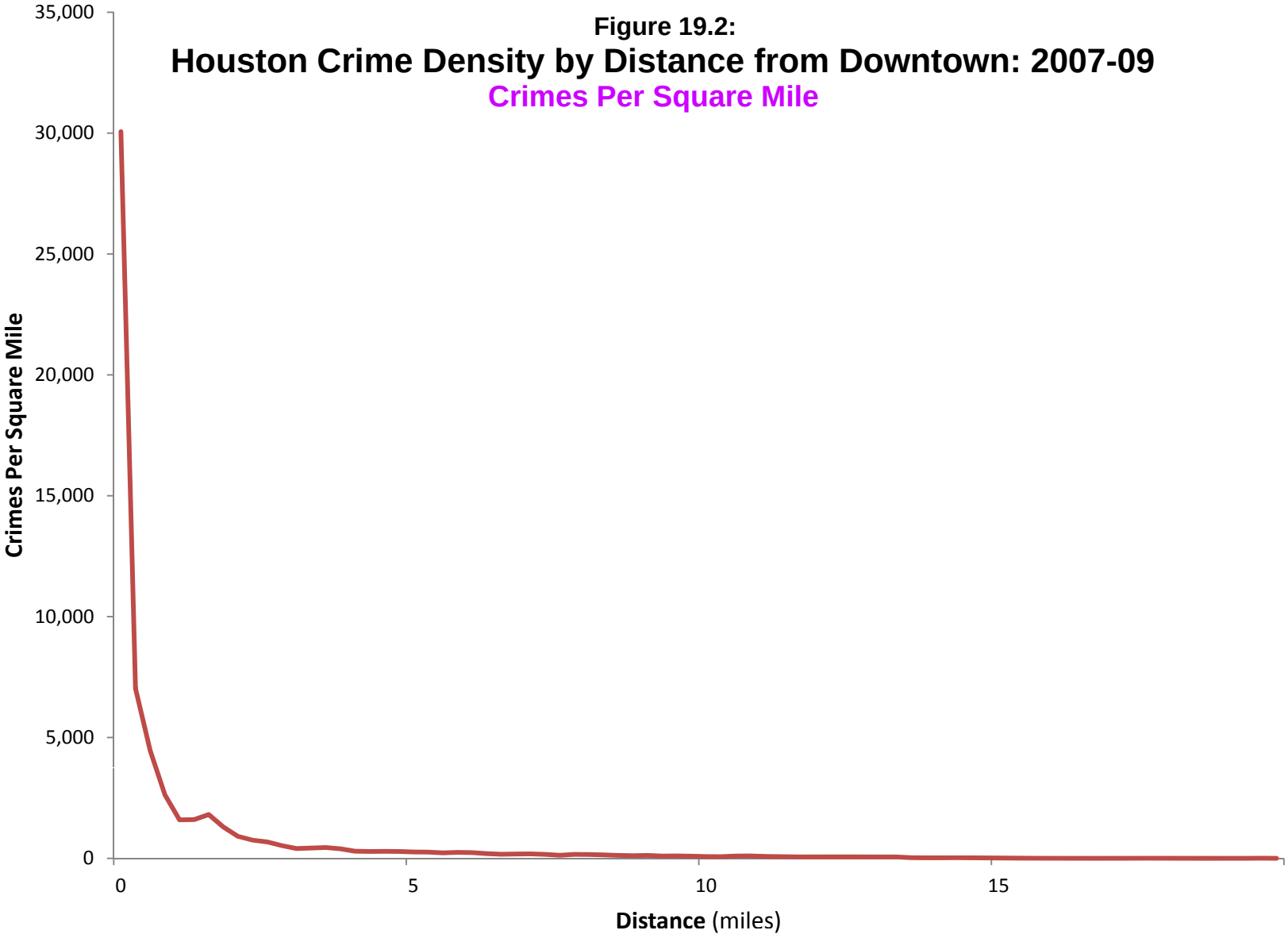


Figure 19.1:
Houston Crimes by Distance from Downtown: 2007-09
Number of Crimes
25,000
20,000
15,000
10,000
5,000
0
Number of crimes
0
5
10
15
20
Distance (miles)

**Figure 19.2:**

## Houston Crime Density by Distance from Downtown: 2007-09

**Crimes Per Square Mile**



that zone's value (of any dependent variable) and the values of nearby zones. This is merely a product of increasing concentration in the central city. Without making the global effect explicit can make it appear that the local effect is stronger than it is.

The disadvantage in building in an explicit global 'distance from center' variable assumes that the relationship will hold in future. The zones may be redefined which might distort the relationship slightly (e.g., what the U.S. Census Bureau does from census to census). Also, urban change may distort the distance relationship over time, for example with decreasing crime rates in central cities (Kneebone & Raphael, 2011). Of course, this applies to any spatial variable and not just to distance from the central area.

Also, the inclusion of a global distance variable may cover up a true local effect. For example, Heskin, Levine and Garrett (2000) examined housing and population change using an OLS model at the edges of four California cities that had rent control with vacancy control provisions.[2] By comparing block groups on both sides of the borders, they were able to show the effect of vacancy control over rents was to reduce the number of rental housing units from 1980 to 1990 in the block groups associated with vacancy control compared to the block groups in cities without vacancy control. Had all the block groups in the compared cities been included (which included two very large cities – Los Angeles and Oakland), the relationship would have been obscured.

In short, there are both global spatial effects and local spatial effects, but they need to be distinguished. To not separate out these effects could easily lead to misinterpretation as the relative importance of local spatial clustering (i.e., local spatial autocorrelation). The ideal would be to include both a global distance variable as well as a test for local spatial autocorrelation. The spatial parameter tests discussed below can do this.

### *Values of nearby zones*

More questionable is the inclusion of values for nearby zones. A number of studies have included the values of nearby zones to account for spatial autocorrelation. For example, Wachter and Cho (1991) showed with an OLS model that the restrictiveness of the zoning in adjacent areas independently increased the price of single family homes in Montgomery County, MD. That is, by including the price of single family home in adjacent communities, they showed that it had an effect on the price of single family homes in the community they were studying; they also included a distance to downtown Washington DC as a statistical control variable.

---

[2] Vacancy control involves maintaining the regulated rent levels even if the unit becomes vacant as opposed to allowing the rent to rise to market levels when a rental unit becomes vacant. At the time, four cities in California had vacancy control provisions.

In a recent study of crime trajectories on street segments between 1993 and 2004 in Seattle, Weisburd, Groff and Yang (2012) used a multinomial logit model to predict eight different crime trajectories (see Chapters 21 and 22 on Discrete Choice Analysis). They used three year intervals to estimate the effects. Among the 28 variables in their model was a spatial lag variable which was the average number of crimes on neighboring street segments within one-quarter of a mile as well as a variable measuring change in spatial lag between the first three year period (1993-95) and the last three year period (2002-04).

The advantage of including the values of nearby zones, no matter how defined, is that it builds in an effect of the nearby zones. It is somewhat intuitive to treat the nearby value as an exogenous variable to a model in that it produces a coefficient that appears to represent the value of those nearby values.

### *Problems with using values of nearby zones*

There are a number of disadvantages with this approach, however. First, treating the values of nearby zones as an independent variable assumes independence of those zones and ignores reciprocal effects. This can cause s*imultaneity bias* (Wikipedia, 2013). That is, the value of the dependent variable in nearby zones is treated as independent of the value of the same variable in the zone being modeled (the central zone). Yet, in reality, the effect is two ways; the central zone influences the values of the nearby zones, and vice versa (i.e., they are interrelated). The result will be that the coefficient will be biased because some of the estimated effect (the coefficient) of the nearby zones is due to the central zone itself (i.e., the value of the central zone is on both sides of the equation). Specifying the values of nearby zones as being independent does not incorporate the simultaneous effect and will almost certainly produce a biased coefficient of the effect.

Second, treating the values of nearby zones as being independent assumes uniformity of their effect throughout the study area. In reality, spatial autocorrelation varies throughout a study area. For example, clustering of events (hot spots) occurs at only some locations, as was discussed in Chapters 7, 8 and 9. By assuming a uniform effect throughout the study area, the variable adds error to the model and may obscure locations where real clustering effects actually occur. The example given above from Heskin, Levine and Garrett (2000) illustrated the very specific local effect of a policy on housing and population change. In practice, any local spatial autocorrelation that affects the value of a central zone will vary throughout a study region, being strong in some places and weak in others. Using a single variable for the values of nearby zones will not capture that specificity.

Third, the grouping of nearby zones into a single measure uses arbitrary weighting of the zones to be included. Either contiguous (adjacent) zones are used or else a relationship is

assumed to operate over a certain distance using a distance decay function (see Chapter 13). The choice of the method for weighting the zones can affect the results substantially. If contiguous zones are used, non-standardized zone size can alter the relationship. For example, in the downtown area of most cities, the zones will be very small, typically a block or two whereas in the suburbs, zones are much larger. Using contiguous zones may not properly cover the spatial autocorrelation effect. In the central city, the effect might extend well beyond one or two blocks whereas in the suburbs, the effect might be smaller than adjacent zones. If distance is used, researcher must make assumptions about the decay function, the type of function used (e.g., linear, negative exponential) as well as the rate of decay. The internal approach to be discussed below also requires the making of assumptions, a point that will be discussed later in the chapter.

Fourth, adding in a spatial autocorrelation variable does not explain the reason for the spatial autocorrelation but simply accounts for some of the additional variance in a model. That is, the spatial autocorrelation variable accounts for additional variance of the dependent variable after all the independent variables have been accounted for. In other words, that there is additional variability that is spatially organized beyond that accounted for by the included independent variables. Note that this criticism applies to an internal spatial parameter as well.

The important thing to realize is that spatial autocorrelation is merely a statistical index created by spatial effects between nearby zones, either clustering or dispersion. It is not a 'thing' or a 'process' but merely a statistical index. The researcher or analyst would do well to find other variables that could explain some of the variability.

### *Eliminating bias from values of nearby zones*

Eliminating the bias from treating the values of nearby zones as exogenous is complicated. There are two main approaches. First, substitute a truly exogenous variable for the externally-defined spatial autocorrelation variable. This is sometimes called an *instrumental* variable (Wikipedia, 2013a). For example, if the number of crimes is the dependent variable and is correlated with alcohol licenses, substituting the number of alcohol licenses in nearby zones for the spatial autocorrelation variable could capture some of the variance associated with nearby zones without adding bias to the estimate.

Second, one could run simultaneous models (i.e., $Y_i$ predicts $Y_j$ as well as $Y_j$ predicts $Y_i$ where $i \neq j$) iteratively many times until the estimates stabilize. In the models discussed below, we use the Markov Chain Monte Carlo (MCMC) approach to produce stabile estimates.

**Internally-estimated Spatial Parameter**

An alternative, and more elegant, approach is to utilize a spatial parameter within the model which is estimated within the calculations themselves. The advantage is that the parameter is estimated simultaneously with the coefficients and will include the reciprocal effects of nearby zones on the central zone, and vice versa. As with a distance-based external variable, the user must make assumptions about the decay of the spatial autocorrelation effect.

There are two common ways to express the internally-estimated spatial parameter, either as a Conditional Autoregressive (CAR) function or as a Simultaneous Autoregressive (SAR) function (De Smith, Goodchild, & Longley, 2007). The CAR function was developed by Besag (1974) while the SAR model was developed by Whittle (1954).

The CAR model is expressed as:

$$E(y_i \mid y_{j \neq i}) = g[\mu_i + \rho \sum_{j \neq i} w_{ij}(y_j - \mu_j)] \tag{19.1}$$

where $g$ is a function relating the expected mean to a linear set of predictors (e.g., Poisson, linear/OLS, logit), $\mu_i$ is the expected value for observation $i$, $w_{ij}$ is a spatial weight between the observation, $i$, and all other observations, $j$ (and for which all weights sum to 1.0), and $\rho$ is a spatial autocorrelation parameter that determines the size and nature of the spatial neighborhood effect. The summation of the spatial weights times the difference between the observed and predicted values is over all other observations ($i \neq j$).

The SAR model has a simpler form and is expressed as:

$$E(y_i \mid y_{j \neq i}) = g[\mu_i + \rho \sum_{j \neq i} w_{ij} y_j] \tag{19.2}$$

where the terms are as defined above. Note, in the CAR model the spatial weights are applied to the difference between the observed and expected values at all other locations whereas in the SAR model, the weights are applied directly to the observed value. In practice, the CAR and SAR models produce very similar results.

In both these cases, the spatial autocorrelation component is estimated simultaneously with the coefficients. That is, the model assumes that the effects of nearby zones on the central zone are reciprocal, each affecting the other. The use of an internal spatial parameter overcomes one of the main problems of incorporating the values of nearby zones. Instead, the spatial parameter is treated as a function of *hyperparameters*, independent parameters that determine its properties.

## MCMC Normal-CAR Model

This is the normal (OLS) model but with a spatial autocorrelation term. For a spatial model, we add a spatial effects parameter, essentially breaking the error term into unexplained variance that is associated with spatial autocorrelation and unexplained variance that has no known associations (i.e., noise).

$$y_i = \mathrm{x}_i^T \beta + \varepsilon_i + \phi_i \tag{19.3}$$

where , $\boldsymbol{\beta}$ is a vector of unknown coefficients for the *k* covariates plus an intercept, $\varepsilon_i$ is the model error independent of all covariates, and $\phi_i$ is a *spatial random effect*, one for each observation. Together, the spatial effects are distributed as a complex *multivariate normal* (or Gausian) density function.

The Normal-CAR model has two mathematical properties. First, both the error term, $\varepsilon_i$, and the dependent variable are normally-distributed. Second, the model incorporates an estimate of local spatial autocorrelation in a CAR format (equation 19.1).

To model the spatial effect, $\phi_i$, we assume the following:

$$p(\phi_i \mid \boldsymbol{\Phi}_{-i}) \propto \exp\left( -\frac{w_{i+}}{2\sigma_\phi^2} \left[ \phi_i - \rho \sum_{j \neq i} \frac{w_{ij}}{w_{i+}} \phi_j \right]^2 \right) \tag{19.4}$$

where $p(\phi_i \mid \boldsymbol{\Phi}_{-i})$ is the probability of a spatial effect given a lagged spatial effect, $w_{i+} = \sum_{i \neq j} w_{ij}$ which sums all over $j$ except $i$ (i.e.,all other zones). This formulation gives a conditional normal density with mean $\rho \sum_{j \neq i} \frac{w_{ij}}{w_{i+}} \phi_j$ and variance $\frac{\sigma_i^2}{w_{i+}}$. The parameter $\rho$ determines the direction and overall magnitude of the spatial effects. The term $w_{ij}$ is a spatial weight function between zones $i$ and $j$ (see below). In the algorithm, the term $\sigma_\phi^2 = 1/\tau_\phi$ and the same variance is used for all observations.

The Phi ($\phi_i$) variable is, in turn, a function of three hyperparameters. The first - Rho ($\rho$), might be considered a global component. The second - Tauphi ($\tau_\phi$), might be considered a local component while the third - Alpha ($\alpha$), might be considered a neighborhood component since it

measures the distance decay. Phi ($\phi_i$) is normally distributed and is a function of Rho and Tauphi.

$$\phi_i \mid \mathbf{\Phi}_{-i} \sim N\left( \rho \sum_{j \neq i}^{n} (w_{ij} / w_{i+}) \phi_j, \sigma_\phi^2 / w_{i+} \right) \tag{19.5}$$

Tauphi, in turn, is assumed to follow a Gamma distribution

$$\tau_\phi = \sigma_\phi^{-2} \sim Gamma(a_\phi, b_\phi) \tag{19.6}$$

where $a_\phi$ and $b_\phi$ are hyper-parameters. For a non-informative prior $a_\phi = 0.01$ and $b_\phi = 0.01$ are used as a default. Since the error term was assumed to be distributed as a Gamma distribution, it is easy to show that $\lambda_i$ follows $Gamma(\psi, \psi e^{-\mathbf{x}_i^T \boldsymbol{\beta} - \phi_i})$. The prior distribution for $\psi$ is again assumed to follow a Gamma distribution

$$\psi \sim Gamma(a_\psi, b_\psi) \tag{19.7}$$

where $a_\psi$ and $b_\psi$ are hyper-parameters. For a non-informative prior $a_\phi = 0.01$ and $b_\phi = 0.01$ are used as a default.

## MCMC Normal-SAR Model

The Normal-SAR model is very similar to the Normal-CAR. The only difference is in the specification of the spatial autocorrelation term. The SAR (or Simultaneous Autoregressive) term is defined as:

$$\phi_i = \rho \sum_{j}^{n} (c_{ij} / c_{i+}) \phi_j + e_i \tag{19.8}$$

where $e_i$ are iid $N(0, \sigma_\phi^2 / c_{i+})$. All the other variables ($c_{ij}, c_{i+}, \rho$) are exactly the same as for the CAR model described above. The Phi ($\phi_i$) variable is estimated using equation 19.5 above.

### Potential Problem in Running MCMC Normal-CAR/SAR Models

Users should be cognizant of a potential problem in using the MCMC Normal model with or without the CAR/SAR spatial autocorrelation parameter. The model is appropriate when the

dependent variable is normally distributed and the CrimeStat MCMC routine will work well under these conditions. However, if the dependent variable is highly skewed, the MCMC Normal often will not produce accurate estimates.

We are not completely sure of the conditions that cause the MCMC Normal to not properly produce a good representation of the data. Users should test whether the MCMC Normal (without the spatial autocorrelation parameter) can replicate the results of the MLE Normal. If it can produce a reasonably close approximation, then the MCMC Normal is converging properly and the results of an MCMC Normal-CAR or MCMC Normal-SAR can be trusted. However, if the MCMC Normal does not produce a reasonably close approximation to the MLE Normal, then the algorithm has not converged properly and the user is advised to use one of the Poisson-based models.

A better convergence can often be obtained by first running the MLE Normal and using the estimated intercept and coefficients as prior values in an MCMC Normal. Chapter 20 discusses the specifics of assigning prior values in an MCMC model.

Also, the MCMC Normal is affected by multicollinearity among independent variables. Because multicollinearity creates ambiguity in the coefficients of the collinear variables, it will affect the stability of the MCMC model. This is also true in MCMC Poisson-based models but has much greater effect for MCMC Normal models. Our advice is to eliminate collinear variables in order that those independent variables in the model are truly independent of each other. This will tend to improve MCMC Normal estimates.

> To repeat, the MCMC Normal is appropriate when the dependent variable is normally-distributed. It is not appropriate for highly skewed dependent variables.

## MCMC Poisson-Gamma-CAR Model

This is the negative binomial model but with a spatial autocorrelation term. Formally, it is defined as:

$$y_i \mid \lambda_i \sim Poisson(\lambda_i) \tag{19.9}$$

with the mean of Poisson-Gamma-CAR organized as:

$$\lambda_i = \exp(\mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i + \phi_i) \tag{19.10}$$

where exp() is an exponential function, $\boldsymbol{\beta}$ is a vector of unknown coefficients for the *k* covariates plus an intercept, and $\varepsilon_i$ is the model error independent of all covariates. The $\xi_i = \exp(\varepsilon_i)$ is assumed to follow the gamma distribution with a mean equal to 1 and a variance equal to $1/\psi$ where $\psi$ is a parameter that is greater than 0, and $\phi_i$ is a *spatial random effect*, one for each observation.

The assumption on the uncorrelated error term $\varepsilon_i$ is the same as in the Poisson-Gamma model. The third term in the expression, $\phi_i$, is a *spatial random effect*, one for each observation. Together, the spatial effects are distributed as a complex *multivariate normal* (or Gausian) density function. In other words, the second model is a spatial regression model within a negative binomial model.

The Poisson-Gamma-CAR model has three mathematical properties. First, the count is Poisson distributed, as is true of all Poisson-based models. Second, the mean is distributed as a Gamma function, similar to the negative binomial model. Third, it incorporates an estimate of local spatial autocorrelation in a CAR format (equation 19.1). The same assumptions about the spatial effect apply for the Poisson-Gamma-CAR model as for the Normal-CAR model.

## MCMC Poisson-Gamma-SAR Model

The Poisson-Gamma-SAR model is very similar to the Poisson-Gamma-CAR. The only difference is in the specification of the spatial autocorrelation term. The SAR (or Simultaneous Autoregressive) term is defined as:

$$\phi_i = \rho \sum_j^n (c_{ij} / c_{i+}) \phi_j + e_i \qquad (19.11)$$

where $e_i$ are iid $N(0, \sigma_\phi^2 / c_{i+})$. All the other variables ($c_{ij}, c_{i+}, \rho$) are exactly the same as for the CAR model described above. The Phi ($\phi_i$) variable is estimated using equation 19.5 above.

## MCMC Poisson-Lognormal-CAR/SAR Model

As described in Chapter 17, the Poisson-Lognormal model has a distribution that is Poisson-distributed.

$$y_i \mid \lambda_i \sim Poisson(\lambda_i) \qquad \text{repeat (16.24)}$$

However, the Poisson mean $\lambda_i$ is organized as:

$$\lambda_i = \exp(\mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i + \phi_i) = \exp(\mathbf{x}_i^T \boldsymbol{\beta} + \phi_i) \cdot \xi_i \tag{19.12}$$

where exp() is an exponential function, $\boldsymbol{\beta}$ is a vector of unknown coefficients for the *k* covariates plus an intercept, $\phi_i$ is the spatial random effect, and $\varepsilon_i$ is the model error independent of all covariates. The error, $\xi_i = \exp(\varepsilon_i)$, is assumed to follow the lognormal distribution with a mean equal to 0 and a variance equal to $\sigma_\varepsilon^{-2} = \tau_\varepsilon \sim Gamma\ (a_\varepsilon, b_\varepsilon)$.

To model the spatial effect, $\phi_i$, equation 19.1 is used for the CAR spatial model while equation 19.2 is used for the SAR spatial model. An application of Poisson-Lognormal-CAR and SAR models is found in Kim and Lim (2010).

## MCMC Binomial Logit-CAR/SAR MODELS

In Chapter 18, we discussed binomial models, the logit and the probit. As with Poisson-based model, these can have a spatial autocorrelation component, too. In CrimeStat, we include a spatial logit model which is the logit model with a spatial autocorrelation term, $\phi_i$.

$$\log\{p_i/(1-p_i)\} = \mathbf{x}_i^T \boldsymbol{\beta} + \phi_i + \varepsilon_i \tag{19.13}$$

The assumption on the uncorrelated error term $\varepsilon_i$ is the same as in the Poisson-Gamma model. The third term in the expression, $\phi_i$, is a *spatial random effect*, one for each observation.

Together, the spatial effects are distributed as a complex *multivariate normal* (or Gausian) density function. In other words, the second model is a spatial regression component within a logit model. The spatial effect, $\phi_i$, is a *spatial random effect*, one for each observation. It can be modeled as either a CAR (equation 19.1) or a SAR (equation 19.2).

## Spatial Weights Function

For all the CAR and SAR models, the spatial weights function, $w_{ij}$, is a function of the neighborhood parameter, $\alpha$, which is a distance decay function. Three distance weight functions are available in *Crimestat*:

**1. Negative Exponential Distance Decay**

$$w_{ij} = e^{-\alpha d_{ij}} \tag{19.14}$$

where $d_{ij}$ is the distance between two zones or points and α is the decay coefficient. The weight decreases with the distance between zones with α indicating the degree of decay.

**2. Restricted Negative Exponential Distance Decay**

$$w_{ij} = Ke^{-\alpha d_{ij}} \tag{19.15}$$

where *K* is 1 if the distance between points is less than equal to a search distance and 0 if it is not. This function stops the decay if the distance is greater than the user-defined search distance (i.e., the weight becomes 0).

**3. Contiguity Function**

$$c_{ij} = w_{ij} \tag{19.16}$$

where $w_{ij}$ is 1 if observation $j$ is within a specified search distance of observation *i (*a neighbor) and 0 if it is not.

## Estimation Procedures for Spatial Models

For each of the spatial regression models used, we follow the same steps that were outlined in Chapter 17. Conceptually, these are:

1. Specifying a functional model and setting up the model parameters.

2. A likelihood function is set up and prior distributions for each parameter are assumed.

3. A joint posterior distribution for all unknown parameters is defined by multiplying the likelihood and the priors.

4. Repeated samples are drawn from this joint posterior distribution.

5. The estimates for all coefficients are based on the results of the *M-L* samples, for example the mean, the standard deviation, the median and various percentiles. Similarly, the overall model fit is based on the *M-L* samples.

**Determining a Distance Decay Function for Alpha**

Each of these steps is applied to the specified models discussed above. For a spatial regression model, a distance function has to be defined. Alpha (α) is the exponent for the distance decay function in the spatial model. Essentially, the distance decay function defines the weight to be applied to the values of nearby records. The weight can be defined by one of three mathematical functions. First, the weight can be defined by a *negative exponential* function,

$$\text{Weight} = e^{-\alpha * d(ij)} \tag{19.17}$$

where *d(ij)* is the distance between observations in specified units (e.g., miles, meters) and *α* is the value for alpha, again consistent with the specified distance units. It is automatically assumed that alpha will be negative whether the user puts in a minus sign or not. The user inputs the alpha value in this box.

Second, the weight can be defined by a *restricted negative exponential* whereby the negative exponential operates up to the specified search distance, whereupon the weight becomes 0 for greater distances

$$\text{Up to Search distance:} \quad \text{Weight} = e^{-\alpha * d(ij)} \qquad \text{for } d(ij) \geq 0,\ d(ij) \leq d_p \tag{19.18}$$

$$\text{Beyond search distance:} \quad 0 \qquad \text{for } d(ij) > d_p \tag{19.19}$$

where $d_p$ is the search distance. The coefficient for the linear component is assumed to be 1.0.

Third, the weight can be defined as a *uniform* value for all other observations within a specified search distance. This is a *contiguity* (or adjacency) measure. Essentially, all other observations have an equal weight within the search distance and 0 if they are greater than the search distance. The user inputs the search distance and units in this box.

**Determining a Reasonable Value for Alpha**

The default function for the weight is a negative exponential with a default alpha value of -1 in miles. For many data sets, this will be a reasonable value. However, for other data sets, it

will not. Reasonable values for alpha with the negative exponential function are obtained with the following procedure:

1. Decide on the measurement units to be used to calculate alpha (miles, kilometers, feet, etc). The default is miles. *CrimeStat* will convert from the units defined for the Primary File input dataset to those specified by the user.

2. Calculate the nearest neighbor distance from the Nna routine on the Distance Analysis I page. These may have to be converted into units that were selected in step 1 above. For example, if the Nearest Neighbor distance is listed as 2000 feet, but the desired units for alpha are miles, convert 2000 feet to miles by dividing the 2000 by 5280.

3. Input the dependent variable as the Z (intensity) variable on the Primary File page.

4. Run the Moran Correlogram routine on this variable on the Spatial Autocorrelation page (under Spatial Description). By looking at the values and the graph, decide whether the distance decay in this variable is very 'sharp' (drops off quickly) or very 'shallow' (drops off slowly).

5. Define the appropriate weight for the nearest neighbor distance:

    a. Assume that the weight for an observation with itself (i.e., distance = 0) is 1.0.

    b. If the distance decay drops off sharply, then a low weight for nearby values should be given. Assume that any observations at the nearest neighbor distance will only have a weight of 0.5 with observations further away being even lower.

    c. If the distance decay drops off more slowly, then a higher weight for nearby values should be given. Assume that any observations at the nearest neighbor distance will have a weight of 0.9 with observations further away being lower but only slightly so.

    d. An intermediate value for the weight is to assume it to be 0.75.

6. A range of alpha values can be solved using these scenarios:

    a. For the sharp decay, alpha is given by:

$$\alpha = \ln(0.5)/NN(distance) \quad (19.20)$$

where NN(distance) is the nearest neighbor distance in specified distance units (e.g., feet, meters, kilometers)

b. For the shallow distance decay, alpha is given by:

$$\alpha = \ln(0.9)/NN(distance) \quad (19.21)$$

where NN(distance) is the nearest neighbor distance.

c. For the intermediate decay, alpha is given by:

$$\alpha = \ln(0.75)/NN(distance) \quad (19.22)$$

where NN(distance) is the nearest neighbor distance.

These calculations will provide a range of appropriate values for α. The diagnostics routine automatically estimates these values as part of its output.

**Value for Zero Distance Between Records**

The advanced options dialogue has a parameter for the minimum distance to be assumed between different records. If two records have the same X and Y coordinates (which could happen if the data are individual events, for example), then the distance between these records will be 0. This could cause unusual calculations in estimating spatial effects. Instead, it is more reliable to assume a slight difference in distance between all records. The default is 0.005 miles but the user can modify this (including substituting 0 for the minimal distance).

**GUIDELINE:**

Note that MCMC spatial regression models will take a very long time to calculate. For large datasets, we recommend using the block sampling method discussed in chapter 17. A rough rule-of-thumb is that if the dataset is larger than 2,000 cases, the block sampling method should be used for spatial MCMC models. Of course, this will depend on the amount of available RAM as well as the processing speed of the computer.

## Examples of Spatial Regression Modeling

### Example 1: MCMC Normal-CAR Analysis of Houston Burglaries

The first example is the MCMC Normal-CAR model using the Houston burglary data set. The data came from the Houston Police Department. There were 26,480 burglaries that occurred in 2006 which were allocated to 1,179 Traffic Analysis Zones (TAZ) within the City of Houston. The independent variables were the number of households in 2006 (estimated by the Houston-Galveston Area Council, the metropolitan planning organization) and the median household income for 2000 (from the 2000 U.S. Census).

Two statistical control variables were used: 1) the area in square miles of the TAZ; and 2) the distance of the zone from downtown Houston. It should be noted that the two control variables are correlated in that TAZs get larger with distance from the center. Still, as they are being used to control for measurement error and global spatial autocorrelation, ambiguity in their interpretation due to this correlation is less critical than that they filter out effects that bias the other variables in the equation.

With a spatial regression model, the user has to provide a value for the distance decay term, alpha (α). The diagnostics routine that was discussed in Chapter 15 provides plausible values of α given the decline in spatial autocorrelation as measured by the Moran Correlogram. The diagnostic calculates the nearest neighbor distance (the average distance of the nearest neighbors for all observations) and then estimates values based on weights assigned to this distance. Three weights are estimated: 0.9, 0.75 and 0.5. We utilized the 0.75 weight. In the example, based on the nearest neighbor distance of 0.45 miles and a weight of 0.75, the α value would be -0.637 for distance units in miles.

Table 19.1 presents the results. For comparison, we ran the MCMC Normal model without the CAR adjustment but with the global statistical control variables (Table 19.2). The R-square of the spatial model is slightly worse than the non-spatial model. But, remember, these are estimates based on samples and will vary from run to run. The log likelihood is better for the non-spatial model than for the spatial model. The AIC and BIC/SC statistics are also better. However, the Mean Absolute Deviation (MAD) and the Mean Squared Predictive Error (MSPE) are similar for the two models. There are subtle differences in the MAD and MSPE between the models (e.g., the MCMC Normal-CAR has better MAD and MSPE scores for the first and third quartiles), but the differences are very small.

In both models, the number of households in the TAZ is positively related to the number of burglaries (i.e., more households, then more burglaries). The median household income is negatively related to the number of burglaries (i.e., burglaries are more likely to occur in poorer

**Table 19.1:**

# Predicting Burglaries in the City of Houston: 2006
# MCMC Normal-CAR Model

(N= 1,179 Traffic Analysis Zones)

| | |
|---|---|
| **DepVar:** | **2006 BURGLARIES** |
| N: | 1,179 |
| Df: | 1,172 |
| Type of regression model: | Poisson with Lognormal dispersion |
| Method of estimation: | MCMC |
| Number of iterations: | 25,000 Burn in: 5,000 |

***Likelihood statistics***

| | |
|---|---|
| Log Likelihood: | -6,369.6 |
| AIC: | 12,593.4 |
| BIC/SC: | 12,753.2 |
| $R^2$: | 0.50 |

***Model error estimates***

| | |
|---|---|
| Mean absolute deviation: | 13.3 |
| 1st (highest) quartile: | 24.2 |
| 2nd quartile: | 11.6 |
| 3rd quartile: | 10.6 |
| 4th (lowest) quartile: | 6.8 |
| Mean squared predicted error: | 481.1 |
| 1st (highest) quartile: | 1,296.1 |
| 2nd quartile: | 320.1 |
| 3rd quartile: | 209.9 |
| 4th (lowest) quartile: | 102.0 |

| Predictor | Mean | Std | t-value[p] | MC error | MC error/ std | G-R stat |
|---|---|---|---|---|---|---|
| **INTERCEPT** | 6.4422 | 0.623 | 10.34*** | 0.012 | 0.018 | 1.002 |
| **HOUSEHOLDS** | 0.0240 | 0.0004 | 56.93*** | 0.000005 | 0.012 | 1.0003 |
| **MEDIAN HOUSEHOLD INCOME** | -0.0002 | 0.00001 | -12.99*** | 0.0000002 | 0.016 | 1.001 |
| *Statistical control* | | | | | | |
| **AREA** | -7.3012 | 0.467 | -15.65*** | 0.006 | 0.012 | 1.001 |
| **DISTANCE TO DOWNTOWN** | 1.8918 | 0.084 | 22.60*** | 0.001 | 0.017 | 1.001 |
| *Spatial autocorrelation* | | | | | | |
| **AVERAGE PHI** | -1.2612 | 0.084 | -14.95 | 0.025 | 0.298 | 3.243 |

*** $p \leq .001$

**Table 19.2:**

# Predicting Burglaries in the City of Houston: 2006

## MCMC Normal Model

(N= 1,179 Traffic Analysis Zones)

| | |
|---|---|
| **DepVar:** | **2006 BURGLARIES** |
| N: | 1,179 |
| Df: | 1,172 |
| Type of regression model: | Poisson with Lognormal dispersion |
| Method of estimation: | MCMC |
| Number of iterations: | 25,000 Burn in: 5,000 |

***Likelihood statistics***

| | |
|---|---|
| Log Likelihood: | -5,303.1 |
| AIC: | 10,615.2 |
| BIC/SC: | 10,618.2 |
| $R^2$: | 0.51 |

***Model error estimates***

| | |
|---|---|
| Mean absolute deviation: | 13.3 |
| 1st (highest) quartile: | 24.9 |
| 2nd quartile: | 10.4 |
| 3rd quartile: | 9.8 |
| 4th (lowest) quartile: | 8.0 |
| Mean squared predicted error: | 472.5 |
| 1st (highest) quartile: | 1,357.7 |
| 2nd quartile: | 262.6 |
| 3rd quartile: | 170.4 |
| 4th (lowest) quartile: | 103.0 |

| Predictor | Mean | Std | t-value[p] | MC error | MC error/ std | G-R stat |
|---|---|---|---|---|---|---|
| **INTERCEPT** | 9.7376 | 1.252 | 7.78*** | 0.022 | 0.016 | 1.002 |
| **HOUSEHOLDS** | 0.0231 | 0.001 | 27.42*** | 0.00001 | 0.011 | 1.0009 |
| **MEDIAN HOUSEHOLD INCOME** | -0.0002 | 0.00002 | -8.82*** | 0.0000004 | 0.015 | 1.001 |
| | *Statistical control* | | | | | |
| **AREA** | -5.7398 | 0.931 | -6.17*** | 0.011 | 0.012 | 1.000 |
| **DISTANCE TO DOWNTOWN** | 1.0556 | 0.167 | 8.88*** | 0.003 | 0.018 | 1.001 |

*** $p \leq .001$

neighborhoods). The two control variables also show consistency in the two models. The area of the TAZ is negatively related to the number of burglaries while distance from downtown Houston is positively related.

As mentioned, these two control variables do correlate negatively with each other (i.e., TAZs in the center of the city are smaller than in the periphery) and one could drop either and get about the same amount of statistical control in the model. The positive association for distance has to do with a number of factors (there are proportionately more residential households farther away from the center whereas there are more mixed land uses in the center; surveillance and police presence is less concentrated farther away; there are a number of high crime, low income neighborhoods distributed quite far from downtown Houston). As mentioned above, distance from the city center frequently shows associations with the number of crimes committed and many other phenomena. It should almost always be included in a regression model.

The biggest differences are in the coefficients. The intercept is smaller for the MCMC Normal-CAR while the coefficients for households and for median household income are about the same while the average Phi coefficient (the average of all the Phi values for individual records) is highly significant. In other words, the spatial autocorrelation component (estimated by Phi) absorbed some of the variance in the dependent variable and 'pulled' this from the intercept. Keep in the mind that the intercept is a constant that is added to the predicted values for all records. In both models, the statistical controls are slightly stronger in the CAR model.

Two points should be noted. First, as mentioned above, the MCMC Normal-CAR (or MCMC Normal-SAR) model assumes that the dependent variable is approximately normally distributed. However, as discussed in Chapter 15, the Houston burglary data by TAZ is highly skewed, and over-dispersed. Thus, the Normal model (whether tested with MLE or MCMC) is not appropriate for a highly skewed dependent variable. The MCMC Normal seems to have done a good job of replicating the MLE Normal for this dataset, but neither model is appropriate for a skewed dependent variable. Instead, one of the Poisson-family models should be used.

Second, the two diagnostic statistics for Phi that indicate whether the distribution has converged on an 'equilibrium' state, namely the MC Error/STD and the G-R statistics, are much higher than would normally be acceptable (i.e., below 0.05 and 1.20 respectively). We are not completely sure why this occurs, but these particular statistics for PHI do not get smaller in the even with a large number of iterations. In this model only, these diagnostic statistics are much higher than with the Poisson models. Users should be aware of this. Most importantly, though, is that users should ensure that the two diagnostic indicators are low for the independent variables, which they are in Table 19.1. This will indicate that the model has converged to equilibrium and can be trusted.

**Example 2: MCMC Poisson-Gamma-CAR Analysis of Houston Burglaries**

In the second example, we ran the Houston burglary data set using a Poisson-Gamma-CAR since this model is appropriate when there is an over-dispersed dependent variable. The procedure we follow is similar to that outlined in Oh, Lyon, Washington, Persaud, and Bared (2003). First, we ran the Poisson-Gamma model for burglaries along with the two statistical control variables (Table 19.3). The associations are similar to the Normal models. The number of households was positively associated with the number of burglaries while the median household income was negatively associated. For the control variables, the area of the TAZ was negatively related (i.e., more burglaries occurred in smaller zones) while the distance of the TAZ from downtown Houston was positively associated (the two control variables are correlated, as mentioned above).

Second, we tested the residual errors for spatial autocorrelation using the Moran's "I" routine in *CrimeStat*. As expected, the "I" for the residuals were highly significant ("I" = 0.0127; $p \leq .0001$) indicating that there is still substantial spatial autocorrelation in the error term in spite of the inclusion of the global spatial autocorrelation term (distance to downtown Houston).

Third, we then ran a Poisson-Gamma-CAR model of the Houston burglaries along with the two statistical control variables – area of the zone and distance to downtown Houston. We utilized an alpha, $\alpha$, value of -0.637 for distance units in miles as in the MCMC Normal-CAR model.

Table 19.4 present the results. The likelihood statistics indicated that the overall model fit was similar to that of the Poisson-Gamma model. The log likelihood was fractionally higher with the Poisson-Gamma-CAR and the DIC, AIC BIC/SC, and Deviance were slightly lower. But, the values are almost identical in the two models. The MAD values are identical while the MSPE value for the non-CAR model is slightly lower than for the Poisson-Gamma-CAR model. But, again, the differences were small.

Regarding individual coefficients, the intercept, the two independent variables, and the two control variables have values similar to that of MCMC Poisson-Gamma. Note, though, that the coefficient value for the intercept is now slightly smaller. The reason is that the spatial effects, the $\phi_i$ values, have absorbed some of the variance that was previously associated with the intercept. The table presents an average Phi value over all observations. The overall average was not statistically significant. However, Phi values for individual coefficients were output as an individual file and the predicted values of the individual cases include the individual Phi values.

**Table 19.3:**
# Predicting Burglaries in the City of Houston: 2006
# MCMC Poisson-Gamma Model
(N= 1,179 Traffic Analysis Zones)

| | | | |
|---|---|---|---|
| **DepVar:** | **2006 BURGLARIES** | | |
| N: | 1179 | | |
| Df: | 1172 | | |
| Type of regression model: | Poisson-Gamma-CAR | | |
| Method of estimation: | MCMC | | |
| Number of iterations: | 25000 | Burn in: 5000 | |
| Distance decay function: | Negative exponential | | |
| ***Likelihood statistics*** | | | |
| Log Likelihood: | -4,345.1 | | |
| DIC: | 8,702.6 | | |
| AIC: | 8,702.2 | BIC/SC: | 8,732.7 |
| Deviance: | 1,390.0 | p-value of deviance: | 0.0001 |
| ***Model error estimates*** | | | |
| Mean absolute deviation: | 37.2 | | |
| Mean squared predicted error: | 50,209.5 | | |
| ***Over-dispersion tests*** | | | |
| Dispersion multiplier: | 1.3 | p-value of dispersion multiplier: | 0.0001 |
| Inverse dispersion multiplier: | 0.8 | | |

| Predictor | Mean | Std | t-value[p] | MC error | MC error/ std | G-R stat |
|---|---|---|---|---|---|---|
| **INTERCEPT** | 1.8663 | 0.083 | 22.45*** | 0.002 | 0.022 | 1.002 |
| **HOUSEHOLDS** | 0.0010 | 0.00007 | 15.62*** | 0.000001 | 0.014 | 1.000 |
| **MEDIAN HOUSEHOLD INCOME** | -0.00001 | 0.000001 | -8.33*** | 0.00000002 | 0.017 | 1.002 |
| | *Statistical control* | | | | | |
| **AREA** | -0.4222 | 0.065 | -6.49*** | 0.001 | 0.016 | 1.002 |
| **DISTANCE TO DOWNTOWN** | 0.1298 | 0.010 | 12.43*** | 0.0002 | 0.020 | 1.004 |

*** p≤.001

**Table 19.4:**

# Predicting Burglaries in the City of Houston: 2006
# MCMC Poisson-Gamma-CAR Model

(N= 1,179 Traffic Analysis Zones)

| | | | |
|---|---|---|---|
| **DepVar:** | **2006 BURGLARIES** | | |
| N: | 1179 | | |
| Df: | 1172 | | |
| Type of regression model: | Poisson-Gamma-CAR | | |
| Method of estimation: | MCMC | | |
| Number of iterations: | 25000 | Burn in: 5000 | |
| Distance decay function: | Negative exponential | | |

***Likelihood statistics***

| | | | |
|---|---|---|---|
| Log Likelihood: | -4,343.4 | | |
| DIC: | 8,694.1 | | |
| AIC: | 8,700.7 | BIC/SC: | 8,736.2 |
| Deviance: | 1,387.0 | p-value of deviance: | 0.0001 |

***Model error estimates***

| | |
|---|---|
| Mean absolute deviation: | 37.2 |
| Mean squared predicted error: | 50,386.2 |

***Over-dispersion tests***

| | | | |
|---|---|---|---|
| Dispersion multiplier: | 1.3 | p-value of dispersion multiplier: | 0.0001 |
| Inverse dispersion multiplier: | 0.8 | | |

| Predictor | Mean | Std | t-value[p] | MC error | MC error/ std | G-R stat |
|---|---|---|---|---|---|---|
| **INTERCEPT** | 1.8642 | 0.083 | 22.34*** | 0.002 | 0.039 | 1.001 |
| **HOUSEHOLDS** | 0.0010 | 0.00007 | 15.71*** | 0.000001 | 0.021 | 1.001 |
| **MEDIAN HOUSEHOLD INCOME** | -0.00001 | 0.000001 | -8.27*** | 0.00000002 | 0.020 | 1.004 |
| *Statistical control* | | | | | | |
| **AREA** | -0.4267 | 0.065 | -6.60*** | 0.001 | 0.012 | 1.001 |
| **DISTANCE TO DOWNTOWN** | 0.1301 | 0.010 | 12.70*** | 0.0002 | 0.018 | 1.001 |
| *Spatial autocorrelation* | | | | | | |
| **AVERAGE PHI** | -0.0003 | 0.002 | -0.17n.s. | 0.00003 | 0.056 | 1.010 |

n.s. Not significant
*** p≤.001

Figure 19.3 shows the residual errors from the Poisson-Gamma-CAR model. As seen, the model overestimated on the west, southwest and southeast parts of Houston. This is in contrast with the normal model which underestimated in the southwest part of Houston with similar overestimation in the west and southeast (not shown). The Poisson-Gamma-CAR model has shifted the estimation errors to the southwest.

When we look at spatial autocorrelation among the residual errors, we now find much less spatial autocorrelation. The Moran's "I" test for the residual errors was 0.0127, virtually identical to that Poisson-Gamma model. It is significant, but much less than with the raw data.

***Spatial Autocorrelation of the Residuals from the Poisson-Gamma-CAR model***

To see the effect of the spatial CAR adjustment, Table 19.5 presents the Moran "I" values and the Getis-Ord "G" values for a search area of 1 mile for the raw dependent variable (2006 burglaries) and four separate models – the MCMC normal, the MCMC Normal-CAR, the MCMC Poisson-Gamma (non-spatial), and the MCMC Poisson-Gamma-CAR, along with the Φ coefficient from the Poisson-Gamma-CAR model.

Moran's "I" tests for positive and negative spatial autocorrelation. A positive value indicates that adjacent zones are similar in value while a negative value indicates that adjacent zones are very different in value (i.e., one being high and one being low). As can be seen, there is positive spatial autocorrelation for the dependent variable and for each of the four comparison models. However, the amount of positive spatial autocorrelation decreases substantially. With the raw variable – the number of 2006 burglaries per zone, there is sizeable positive spatial autocorrelation. However, the models reduce this substantially by accounting for some of the variance of this variable through the two independent variables and through the two control variables (area of the TAZ and distance from downtown Houston). The CAR adjustment reduces the spatial autocorrelation for the Normal model, but the two negative binomial (Poisson-Gamma) models have the same amount.

The Getis-Ord "G" statistic, however, distinguishes two types of positive spatial autocorrelation, positive spatial autocorrelation where the zones with high values are adjacent to zones also with high values (high positive) and positive spatial autocorrelation where the zones with low values are adjacent zones also with low values (low positive). This is a property that Moran's "I" test cannot do.

The "G" has to be compared to an expected "G", which is essentially the sum of the weights. However, when used with negative numbers, such as residual errors, the "G" has to be

**Figure 19.3:**

## Predicting Burglaries in the City of Houston: 2006

### Residual Errors from Poisson-Gamma-CAR Model

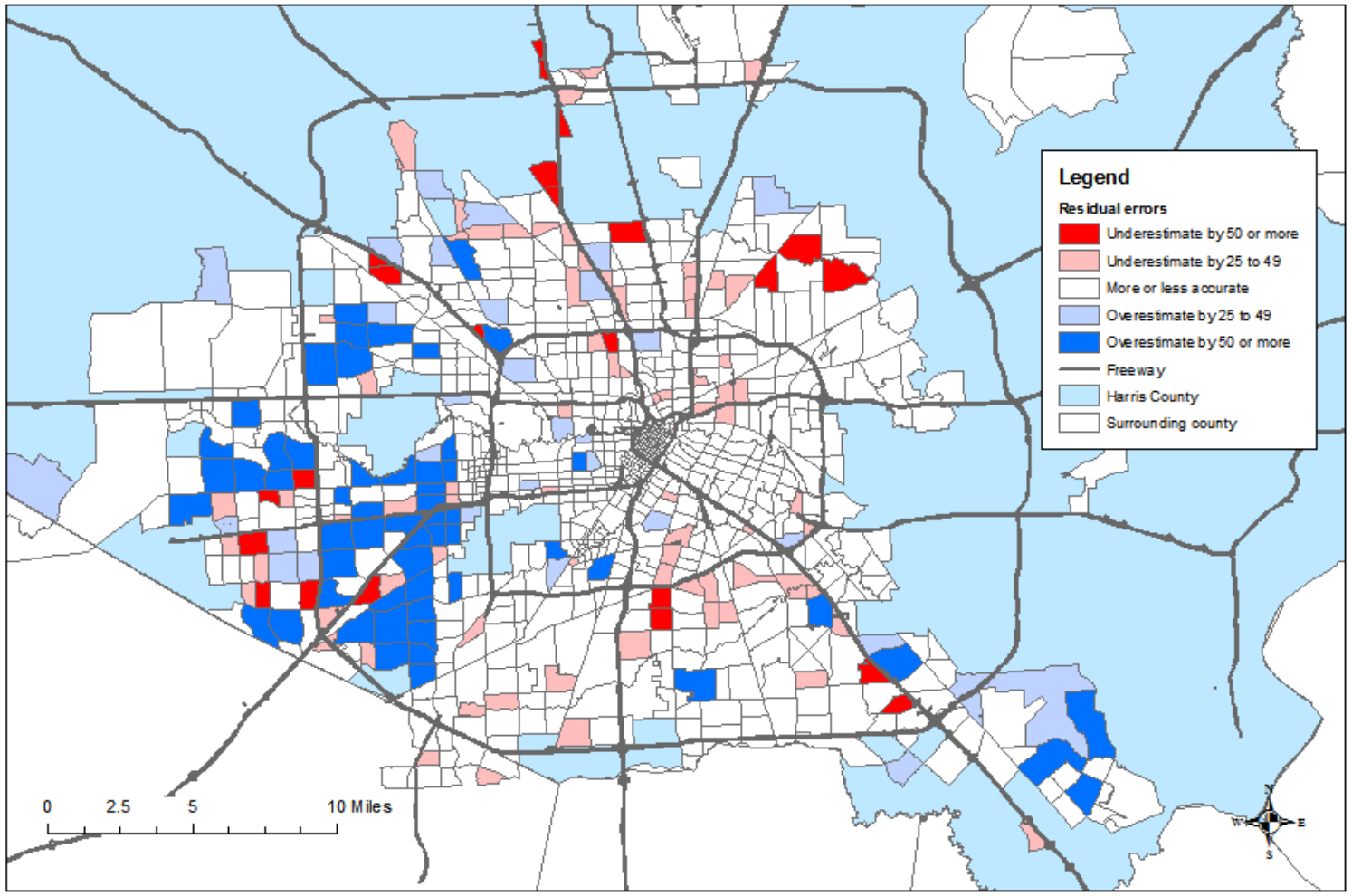


**Table 19.5:**
## Spatial Autocorrelation in Residual Errors of the Houston Burglary Model Comparing Different Poisson Models

| | | Residual Errors | | | | |
|---|---|---|---|---|---|---|
| | **Raw Dependent Variable** | **Normal Model** | **Normal-CAR Model** | **MCMC Poisson-Gamma Model** | **Poisson-Gamma-CAR Model** | **Poisson Gamma-CAR Φ Coefficient** |
| **Moran's "I"** | 0.252**** | 0.022**** | 0.011**** | 0.013**** | 0.013**** | 0.004n.s. |
| **Getis-Ord "G"** (1 mile search radius) | 0.007**** | -1.989**** | 0.618* | 0.013n.s. | 0.013n.s. | 0.081n.s. |

n.s. Not significant
* p≤.05
**** p≤.0001

compared with a simulation envelope. The statistical test for "G" in Table 19.5 tested whether the observed "G" was higher than the 97.5th or 99.5th percentiles (high positive) or lower than the 2.5th or 0.5th percentiles (low positive) of the simulation envelope.

The results show that the "G" for the raw burglary values was *high positive*, meaning that zones with many burglaries tended to be near other zones also with many burglaries. For the analysis of the residual errors, however, the "G" for the Normal model was negative and significant, meaning that it shows negative spatial autocorrelation. That is, the clustering occurs because zones with high residual errors are predominately near other zones with low residual errors. The Normal-CAR model has a "G" that is positive and significant, meaning that it has done the opposite (zones with high residual errors are predominately near to other zones with high residual errors).

On the other hand, the residuals errors for the MCMC Poisson-Gamma and for the MCMC Poisson-Gamma-CAR models are not significant. In other words, these models have accounted for much of the effect measured by the "G" statistic.

The last column analyzes the spatial autocorrelation tests on the individual Phi coefficients. There is spatial autocorrelation for the Phi values, as seen by a very significant Moran "I" value, but it is neither a 'high positive' or a 'low positive' based on the "G" test. In other words, the Phi values appear to be neutral with respect to the clustering of residual errors.

Figure 19.4 shows the distribution of the Phi values. By and large, the spatial adjustment is very minor in most parts of Houston with its greatest impact at the edges, where one might expect some spatial autocorrelation due to very low numbers of burglaries and 'edge effects'.

Putting this in perspective, the spatial effects in the Poisson-Gamma-CAR model are small adjustments to the predicted values of the dependent variable. They slightly improve the predictability of the model but do not fundamentally alter it. Keep in mind that spatial autocorrelation is a statistical effect of some other variable operating that is not being measured in the model. Spatial autocorrelation is not a 'thing' or a process but the result of not adequately accounting for the dependent variable.

In theory, with a correctly specified model, the variance of the dependent variable should be completely explained by the independent variables with the error term truly representing random error. Thus, there should be no spatial autocorrelation in the residual errors under this ideal situation. The example that we have been using is an overly simple one. There are clearly other variables that explain the number of burglaries in a zone other than the number of households and the median household income – the types of buildings in the zone, the street layout, lack of visibility, the types of opportunities for burglars, the amount of surveillance, and so forth. The existence of a spatial effect is an indicator that the model could still be improved by adding more variables.

**Example 3: Modeling Burglary Risk in Houston**

In Chapter 17, we examined risk analysis and used the MCMC algorithm with the Poisson-Gamma model to estimate risk. This can be extended to spatial analysis. To illustrate this type of model, we ran an MCMC Poisson-Gamma-CAR model on the Houston burglary data using the number of households as the exposure variable. There was, therefore, only one independent variable, median household income. Table 19.6 shows the results along with the expanded output that is obtained by clicking on the 'Expanded output' button.

The summary statistics indicate that the overall model fit is good. The log likelihood is high while the AIC and BIC are moderately low. Compared to the non-exposure burglary model (Table 19.5), the model does not fit the data as well. The log likelihood is lower (i.e., more negative) while the AIC and BIC are higher. Further, the DIC is very high

For the model error estimates, the MAD and the MSPE are smaller, suggesting that the burglary risk model is more precise, though not more accurate. However, the dispersion statistics indicate that there is ambiguity over-dispersion. The dispersion multiplier is very low which, by itself, would suggest that a "pure" Poisson model could be used.

**Figure 19.4:**

# Predicting Burglaries in the City of Houston: 2006

## Phi Coefficients from Poisson-Gamma-CAR Model

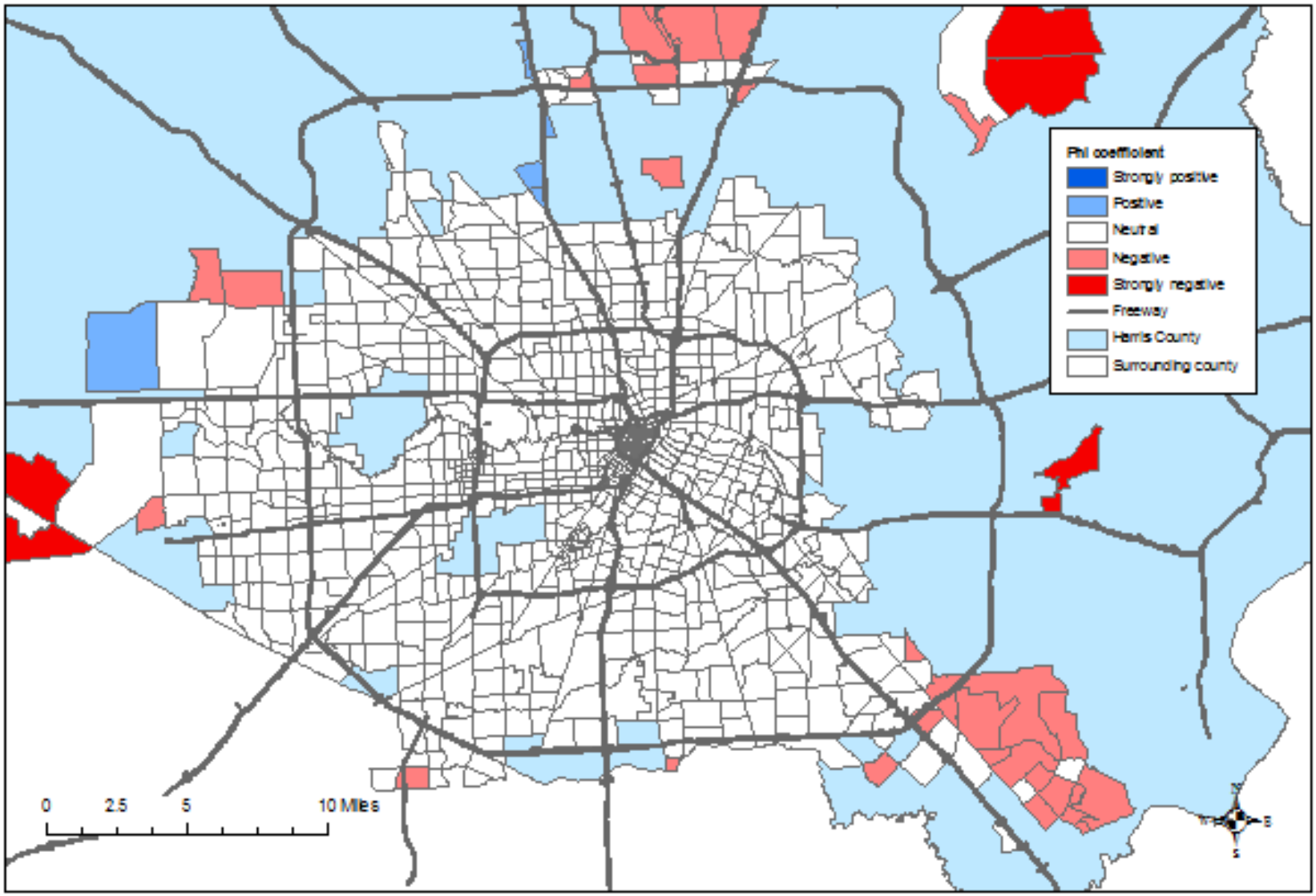


**Table 19.6:**
**Predicting Burglary Risk in the City of Houston: 2006**
**MCMC Poisson-Gamma-CAR Model with Exposure Variable**
**Extended Output**
(N= 1,179 Traffic Analysis Zones)

| | | | |
|---|---|---|---|
| **DepVar:** | **2006 BURGLARIES** | | |
| N: | 1,179 | | |
| Df: | 1,173 | | |
| Type of regression model: | Poisson-Gamma-CAR | | |
| Method of estimation: | MCMC | | |
| Number of iterations: | 25000 | | |
| Burn in: | 5000 | | |
| Distance decay function: | Negative exponential | | |
| Likelihood statistics | | | |
| Log Likelihood: | -3,743.7 | | |
| DIC: | 187,190.2 | | |
| AIC: | 7,499.4 | BIC/SC: | 7,529.9 |
| Deviance: | 996.8 | p-value of deviance: | 0.0001 |
| Model error estimates | | | |
| Mean absolute deviation: | 8.4 | | |
| Mean squared predicted error: | 198.9 | | |
| Over-dispersion tests | | | |
| Dispersion multiplier: | 0.5 | p-value of dispersion multiplier: | n.s. |
| Inverse dispersion multiplier: | 1.8 | | |

| Predictor | Mean | Std | t-value[p] | MC error | MC error/ std | G-R stat |
|---|---|---|---|---|---|---|
| Exposure/offset variable: | | | | | | |
| **HOUSEHOLDS** | 1.0 | | | | | |
| Linear predictors: | | | | | | |
| **INTERCEPT** | -2.0787 | 0.088 | -28.55*** | 0.003 | 0.036 | 1.007 |
| **MEDIAN HOUSEHOLD INCOME** | -0.00002 | 0.000001 | -11.49*** | 0.00000004 | 0.028 | 1.005 |
| *Statistical control* | | | | | | |
| **AREA** | -1.0326 | 0.113 | -9.15*** | 0.005 | 0.046 | 1.011 |
| **DISTANCE TO DOWNTOWN** | 0.0391 | 0.013 | 3.02** | 0.001 | 0.043 | 1.004 |
| *Spatial autocorrelation* | | | | | | |
| **AVERAGE PHI** | 0.0536 | 0.0320 | 1.74n.s. | 0.001 | 0.025 | 1.001 |

n.s. Not significant ** p≤.01 *** p≤.001

**Table 19.6: (continued)**

| Percentiles | 0.5th | 2.5th | 97.5th | 99.5th\ |
|---|---|---|---|---|
| **INTERCEPT** | -2.3029 | -2.2505 | -1.9047 | -1.8500 |
| **MEDIAN HOUSEHOLD INCOME** | -0.00002 | -0.00002 | -0.00001 | -0.00001 |
| **AREA** | -1.3275 | -1.2555 | -0.8160 | -0.7480 |
| **DISTANCE TO DOWNTOWN** | 0.0060 | 0.0142 | 0.0650 | 0.0738 |
| **AVERAGE PHI** | -0.0251 | -0.0053 | 0.1164 | 0.1410 |

Looking at the coefficients, the offset variable (number of households) has a coefficient of 1.0 because it is defined as such. The coefficient for median household income is still negative, but is stronger than in Table 19.5. The effect of standardizing households as the baseline exposure variable has increased the importance of household income in predicting the number of burglaries, controlling for the number of households. Finally, the average Φ value is positive but not significant, similar to what it was in Table 19.5.

***Expanded output***

The use of t-tests to evaluate whether coefficients are significantly different than zero depends on whether the underlying distribution for the coefficients is normal or not. In the case of skewed count data and complex models that are the products of multiple individual functions, it is not always clear whether that assumption is valid or not. Consequently, the *CrimeStat* MCMC module allows the output of statistics that show the distribution of the coefficients in terms of several percentiles: 0.5%, 2.5%, 5%, 10%, 25%, 50%, 75%, 90%, 95%, 97.5% and 99.5%. To obtain these percentiles, the user simply checks the 'Expanded output' box on the MCMC interfact.

In table 19.6 above, we have shown only four of them, the 0.5$^{th}$, 2.5$^{th}$, 97.5$^{th}$, and 99.5$^{th}$ percentiles. The 2.5$^{th}$ and 97.5$^{th}$ represent 95% credible intervals for a two-tailed test while the 0.5$^{th}$ and 99.5$^{th}$ represent 99% credible intervals also for a two-tailed test.

One way to interpret the percentiles is to check whether a coefficient of 0 (the 'null hypothesis') or any other particular value falls outside the 95% or 99% credible intervals. For example, with the intercept term, the 95% credible interval is defined by -2.4365 to -2.1292. Since both are negative, clearly a coefficient of 0 is outside this range; in fact, it is outside the 99% credible interval as well (-2.4879 to -2.0810). In other words, the intercept is *significantly* different than 0, though the use of the term 'significant' is different than with the usual

asymptotic normality assumptions since it is based on the distribution of the parameter values from the MCMC simulation.

Of the other parameters that were estimated, median household income is also significant beyond the 99% credible interval but the Φ coefficient is not significantly different than a 0 coefficient (i.e., a Φ of 0 falls between the 2.5$^{th}$ and the 97.5$^{th}$ percentiles).

In other words, percentiles can be used as a non-parametric alternative to the t- or Z-test. Without making assumptions about the theoretical distribution of the parameter value (which the t- and Z-test do – they are assumed to be normal or near normal for "t"), significance can be assessed empirically. Usually, the t-test and the percentile distribution will lead to the same inference, which they do in table 19.6. But, it is possible that they could differ.

In summary, in risk analysis, an exposure variable is defined and held constant in the model. Thus, the model is really a risk or rate model that relates the dependent variable to the baseline exposure. The independent variables are now predicting the rate, rather than the count by itself.

**Example 4: MCMC Binomial Logit-CAR Analysis of Houston Robberies**

A final example of spatial regression modeling applies the spatial autocorrelation component to the binomial logit model. The test is whether weapons were used in 3,709 Houston robberies that occurred in 2007-09 in which the offender had been arrested. This was the example presented in Chapter 18. The dependent variable was whether a physical weapon had been used, either a firearm, a knife, a stick or another physical object, compared to physical force or the threat of force. The independent variables were the age and gender of the offender, the number of suspects involved, whether the robbery occurred at night (6 PM – 6 AM), the median household income of the zone in which the robbery occurred, and the distance between the robbery location and downtown Houston.

However, now we will examine the distribution using a spatial autocorrelation function, the conditional autoregressive function. The diagnostic routine was run in CrimeStat to determine an appropriate distance decay value (α); this turned out to be -0.4237 in miles. Table 19.7 presents the results.

For this model, the block sampling method discussed in Chapter 17 was used. Comparing these results with the non-spatial binomial logit model for Houston robbery weapon use (Table 18.2), the log-likelihood, AIC and BIC values are slightly stronger for the spatial model than the non-spatial model, suggesting that the spatial adjustment to individual records has improved the overall probability. The goodness of fit statistics (the mean absolute deviation and mean squared

**Table 19.7:**

# Weapon Use by 2007-09 Houston Robbers: MCMC Binomial Logit-CAR Model

(N=3,709 Robberies with Known Origin & Destination Coordinates)

| | | | |
|---|---|---|---|
| **DepVar:** | **WEAPON USE IN ROBBERY** | | |
| N: | 3,709 | | |
| Df: | 3,700 | | |
| Type of regression model: | Logit | | |
| Number of block samples: | 25 | Average block sample size: 395.2 | |
| Method of estimation: | MCMC | | |
| Number of iterations: | 25,000 | Burn in: 5,000 | |
| ***Likelihood statistics*** | | | |
| Log Likelihood: | -2,501.3 | | |
| AIC: | 5,020.5 | | |
| BIC/SC: | 5,076.5 | | |
| Deviance: | -1,204.4 | p: 0.0001 | |
| Pearson Chi-square: | 1,359.9 | p: 0.0001 | |
| ***Model error estimates*** | | | |
| Mean absolute deviation: | 0.4 | | |
| 1st (highest) quartile: | 0.3 | | |
| 2nd quartile: | 0.3 | | |
| 3rd quartile: | 0.4 | | |
| 4th (lowest) quartile: | 0.6 | | |
| Mean squared predicted error: | 0.2 | | |
| 1st (highest) quartile: | 0.1 | | |
| 2nd quartile: | 0.1 | | |
| 3rd quartile: | 0.3 | | |
| 4th (lowest) quartile: | 0.4 | | |
| ***Dispersion tests*** | | | |
| Adjusted deviance: | -0.3 | p: n.s. | |
| Adjusted Pearson Chi-Square: | 0.4 | p: n.s. | |

| Predictor | Mean | Adj. Std | Adj. t-value[p] | . | MC error | MC error/ std | G-R stat | Odds ratio |
|---|---|---|---|---|---|---|---|---|
| Intercept: | 0.6784 | 0.495 | 4.20*** | | 0.019 | 0.038 | 1.011 | 1.971 |
| **AGE** | -0.0242 | 0.012 | -6.27*** | | 0.0004 | 0.032 | 1.011 | 0.976 |
| **GENDER** | -0.6122 | 0.372 | -5.05*** | | 0.005 | 0.013 | 1.003 | 0.542 |
| **# SUSPECTS** | 0.3486 | 0.147 | 7.25*** | | 0.004 | 0.025 | 1.006 | 1.417 |
| **NIGHT** | 0.6753 | 0.323 | 6.40*** | | 0.005 | 0.015 | 1.005 | 1.965 |
| **MED HH INC** | -0.000006 | 0.0000 | -2.21* | | 0.000 | 0.026 | 1.004 | 1.000 |
| **DISTANCE TO DOWNTOWN** | 0.0384 | 0.023 | 5.04*** | | 0.000 | 0.016 | 1.004 | 1.039 |
| **AVERAGE PHI** | -0.0006 | 0.005 | -0.33n.s. | | 0.000 | 0.016 | 1.006 | 0.999 |

*** p≤.001 ** p≤.01 * p≤.05 n.s. Not significant

predicted error) are slightly lower for the spatial model than for the non-spatial model. In particular, the second and third quartiles show slightly lower errors for the Mean Absolute Deviation in the spatial model than the non-spatial model.

The coefficients and adjusted standard errors are very similar between the two models. They differ only in the second or third decimal place. The biggest difference is for nighttime weapon use, where the spatial coefficient is 0.6753 compared to the non-spatial coefficient of 0.5249. This suggest that when spatial location is considered, the nighttime effect in serious weapon use is actually stronger; that is, because the bulk of robberies occur during the daytime but are more clustered spatially, the use of weapons during robberies actually increases at nighttime when controlling for spatial location.

As with the Poisson-Gamma-CAR model, the overall spatial autocorrelation coefficient (Average Phi) is not significant. This is not surprising since the CAR spatial adjustment is done for individual records. In short, the binomial logit-CAR model has produced a slightly better fit to the data than the non-spatial binomial logit model. For prediction, one would use the spatial version because of its better fit.

## Caveat

As mentioned earlier, any spatial regression model is attempting to identify a spatial effect due to clustering, dispersion or some combination whereby the values of nearby zones are similar or different than the central zone (the zone being modeled). In effect, the error term of the model is broken into two parts, one associated with a spatial effect (most likely clustering of nearby zones but sometimes dispersion – negative spatial autocorrelation) and the other with unexplained variance.

What this usually signifies is that there are missing variables that should be included in the model, but which are not. For example, Levine (1999) examined the effects of local growth control measures on housing production in California counties and cities using an OLS spatial lag model (Anselin, 1992). The initial model showed a significant negative spatial effect. However, it was discovered that this was mostly the result of low population density. When density was added to the model, the negative spatial lag effect disappeared.

The important thing to realize with these models is that they identify some variability associated with the dependent variable that needs to be explained. The spatial effect is not real, but merely a statistical artifact of examining similarities or differences between nearby zones in the dependent variable. The spatial indices are useful in that they will indicate whether there is a general spatial effect covering all observations (e.g., distance from downtown; area of the zone) or whether clustering or dispersion is specific to only a limited number of observations (e.g., the

$\phi_i$ coefficient). However, ultimately, the researcher needs to find other variables that account for these effects in order to produce a more stable and realistic model.

## Summary

To summarize, in this chapter we have gone through a number of spatial regression models that apply to normal, Poisson-distributed and binomial logit models. The choice of any of these models is going to depend on the actual distribution of the dependent variable and the underlying assumptions for that model. For example, an MCMC Normal-CAR or MCMC Normal-SAR model only applies if the dependent variable is normally distributed; the use of such a model with non-normal data will usually lead to biased coefficient estimates. Similarly, the two Poisson spatial regression models presented, the MCMC Poisson-Gamma-CAR/.SAR and the MCMC Poisson-Lognormal-CAR/SAR, are applicable if the dependent variable is a count variable or is highly skewed with an absolute zero minimum. The difference is that the MCMC Poisson-Lognormal-CAR/SAR model is used when there is a small sample and a low sample mean (i.e., most zones have 0 events). Finally, the MCMC Binomial Logit-CAR/SAR model is applicable when the dependent variable is binomial and takes the value 0 or 1.

For each of these, the user must define an appropriate distance decay function (α) on the Advanced Options page of the Regression I module. On the interface of the Regression I module, the user can check the diagnostics box to provide plausible values of α based on a Moran Correlogram (see Chapter 5).

In each of these cases, though, the user is advised to first fit a non-spatial model to see if it produces meaningful results. There are two reasons for this. First, unless the independent variables are properly chosen, ambiguity can be introduced by adding a spatial parameter since it is capturing unobserved variability. By 'properly', we mean that all the independent variables are relatively independent (i.e., little multicollinearity) and statistically significant. If a *clean* non-spatial model can be developed first, then adding a spatial autocorrelation component will allow the user to see whether there is clustering among the observations that could account for some of the effects assigned to the independent variables. But, if the model is not clean, then the results are liable to be confusing.

The second reason is practical. The spatial regression models can take a long time to run, as much as several hours. It is more practical to develop a non-spatial model before trying to fit a spatial to it. Almost always, a distance to the nearest city center should be included to capture global spatial effects. Otherwise, there are liable to be associated with the local effects identified by the CAR or SAR models. A CAR or SAR model should be the last step in the modeling, not the first one.

Finally, Chapter 20 will present an overview of the *CrimeStat* regression module. It should be seen as a guide to running the routines.

# References


Alonso, W. (1964). *Location and Land Use: Towards a General Theory of Land Rent*. Harvard University Press: Cambride, MA.

Anselin, L.. (1992). *SpaceStat: A Program for the Statistical Analysis of Spatial Data*. Santa Barbara, CA: National Center for Geographic Information and Analysis, University of California.

Besag, J., 1974. Spatial interaction and the statistical analysis of lattice systems (with discussion). *J. Roy.Statist. Soc. Ser. B 36*, 192–236.

De Smith, M., Goodchild, M. F., & Longley, P. A. (2007). *Geospatial Analysis* (second edition). Matador: Leicester, U.K.

Heskin, A., Levine, N. & Garrett, M. (2000). "Rent control and vacancy control: a spatial analysis of four California cities". *Journal of the American Planning Association*. 66 (2), 162-176.

Hipp, J. R. (2007). Block, Tract, and Levels of Aggregation: Neighborhood Structure and Crime and Disorder as a Case in Point. *American Sociological Review* 72:659-680.

Kim, H. & Lim, H. J. (2010). Comparison of Bayesian spatio-temporal models for chronic diseases. *Journal of Data Sciences*, 8, 189-211.

Kneebone, E. & Raphael, S. (2011). *City and Suburban Crime Trends in Metropolitan America*. Metropolitan Opportunity Series, Metropolitan Policy Program, Brookings Institution: Washington, DC. http://www.brookings.edu/papers/2011/0526_metropolitan_crime_kneebone_raphael.aspx. Accessed April 28, 2012.

Lam, N. S. & De Cola, L. (1993). *Fractals in Geography*. The Blackburn Press: Caldwell, NJ.

Levine, N. (2011). "Spatial variation in motor vehicle crashes by gender in the Houston Metropolitan Area". *Proceedings of the 4th International Conference on Women's Issues in Transportation. Volume II: Technical Papers*, Transportation Research Board: Washington, DC. 12-25. http://onlinepubs.trb.org/onlinepubs/conf/cp46v2.pdf.

Levine, N. (1999). The effects of local growth management on regional housing production and population redistribution in California, *Urban Studies*. 1999. 36 12, 2047-2068.

# References (continued)


Levine, N. & Lee, P. (2013). Crime travel of offenders by gender and age in Manchester, England. Leitner, M. (ed), *Crime Modeling and Mapping Using Geospatial Technologies,* Springer. 145-178.

Miaou, S. P. (2006). "Coding instructions for the spatial regression models in CrimeStat". Unpublished manuscript. College Station, TX.

Oh, J., Lyon, C., Washington, S., Persaud, B., & Bared , J. (2003). "Validation of FHWA crash models for rural intersections: lessons learned". *Transportation Research Record 1840*, 41-49.

Openshaw, S. (1984). *The Modifiable Areal Unit Problem.* Norwich: Geo Books. ISBN 0-86094-134-5.

von Thünen, J. (1826). *The Isolated State in Relation to Agriculture and Political Economy*. English edition, van Suntum, Ulrich. Palgrave Macmillan:Houndsmills, Basingstoke, Hampshire, England, 2009.

Wachter, S. M. & Cho, M. (1991). "Interjurisdictional price effects of land use controls". *Washington University Journal of Urban and Contemporary Law*, 40, 49-63.

Weisburd, D., Groff, E. R., & Yang, S-M (2012). *The Criminology of Place*. Oxford University Press: New York.

Whittle, P., 1954. On stationary process in the plane. *Biometrika*, 41, 434–449.

Wikipedia (2013a). Instrumental variable. Wikipedia. http://en.wikipedia.org/wiki/Instrumental_variable. Accessed January 31, 2013.

Wikipedia (2013b). Specification (regression). Wikipedia. http://en.wikipedia.org/wiki/Specification_(regression). Accessed January 31, 2013.

Wikipedia (2012). Modifiable Area Unit Problem. Wikipedia. http://en.wikipedia.org/wiki/Modifiable_areal_unit_problem. Accessed May 7, 2012.

Wooldridge, J. (2002). Examining the (Ir)Relevance of Aggregation Bias for Multilevel Studies of Neighborhoods and Crime with an Example Comparing Census Tracts to Official Neighborhoods in Cincinnati. *Criminology* 40:681-710.